JEE Journal of Ecological Engineering

Journal of Ecological Engineering 2022, 23(5), 216–225
https://doi.org/10.12911/22998993/147322
ISSN 2299–8993, License CC-BY 4.0

Received: 2022.02.19
Accepted: 2022.03.15
Published: 2022.04.01

Missing Precipitation Data Estimation Using Long Short-Term Memory Deep Neural Networks

Salim Djerbouai[1]

[1] CEHSD Laboratory, Hydraulics Department, University of M'sila, Ichebila, P.O. Box 166, 28000 M'sila, Algeria
E-mail: salim.djerbouai@univ-msila.dz

ABSTRACT

Due to the spatiotemporal variability of precipitation and the complexity of physical processes involved, missing precipitation data estimation remains as a significant problem. Algeria, like other countries in the world, is affected by this problem. In the present paper, Long Short-Term Memory (LSTM) deep neural Networks model was tested to estimate missing monthly precipitation data. The application was presented for the K'sob basin, Algeria. In the present paper, the optimal architecture of LSTM model was adjusted by trial-and-error-procedure. The LSTM model was compared with the most widely used classical methods including inverse distance weighting method (IDWM) and the coefficient of correlation weighting method (CCWM). Finally, it was concluded that the LSTM model performed better than the other methods.

Keywords: hodna, K'sob basin, missing precipitation data, long short-term memory, CCWM, IDWM.

INTRODUCTION

Precipitation constitutes the most important input data for all type of hydrological modelling. However, in practice, records of precipitation are related to the problem of missing data. Processing the rainfall data with missing observations is a serious problem. Estimating missing precipitation data approaches can range from the simplest weighting methods to artificial intelligence-based approaches. Spatial interpolation methods such as the inverse-distance weighting [Shepard, 1968] and normal-ratio methods [Paulhus and Kohler, 1952] have been used for the estimation of missing precipitation data. The inverse distance weighting method was used in many studies [Wei and McGuinness,1973; Simonton and Osborn, 1980; Garcia et al., 2008; Hurtado et al., 2021]. Several improvements to weighting methods were proposed by Teegavarapu and Chandramouli [2005] and also introduced the coefficient of correlation weighting method. Suhaila et al. [2008] introduced several improvements to the inverse distance and normal ratio methods. The results indicate that the performance of these modified methods improved the estimation of missing precipitation data. Teegavarapu [2019] suggested probability space-based weighting methods to estimate the missing daily precipitation data in Kentucky, USA. This proposed new probability space-based distances constitute conceptually superior alternatives to Euclidean distances. Another spatial interpolation used method is based on Kriging [Ly et al., 2011; Teegavarapu and Chandramouli, 2005; Teegavarapu 2007, Xu et al., 2015; Hurtado et al., 2021].

Aguilera et al. [2020] compared three techniques for missing daily precipitation data estimation; spatio-temporal Kriging, multiple imputation by chained equations through predictive mean matching, and the random forest algorithm in the Almonte Marismas aquifer in Spain. They found that the spatio-temporal Kriging is the most robust method. Teegavarapu [2012] compared new mathematical programming models with other techniques (multiple linear regression, nonlinear least-square optimization, Kriging, global and local trend surface as well as thin-plate spline models) for estimating the missing daily precipitation data in the state of Kentucky, USA. The results indicated that the proposed new mathematical programming formulations are superior to those obtained from all the other used methods. Bárdossy and Pegram [2014] compared a new copula-based method with other techniques (regression, Kriging, multiple linear regression) for infilling missing daily and monthly rain gauge data, in the Southern Cape region of South Africa, and the results indicated that the copula-based methods are superior to the others.

Recently, artificial intelligence-based approaches have been widely used in the field of hydrology. Imputations of precipitation data using artificial intelligence-based approaches were reported by multiple studies. Teegavarapu et al. [2009] proposed a fixed functional set genetic algorithm method (FFSGAM). The method uses genetic algorithms and a nonlinear optimization formulation to obtain optimal functional forms to estimate the missing daily precipitation data in the state of Kentucky, USA. They found that the proposed method performed better than traditional inverse distance weighting technique. Random forest approaches were proposed by Mital et al. [2020], to estimate daily precipitation records, in the Upper Colorado water resource region-USA. They found the correlation between references and target stations that influences the performance of the proposed model. Artificial neural networks (ANNs) were applicated in many studies [Kajornrit et al., 2012; Teegavarapu and Chandramouli, 2005; Coulibaly and Evora, 2007; Teegavarapu,2007; Kim and Pachepsky, 2010; Hasanpour Kashani and Dinpashoh, 2012; Londhe et al., 2015 Teegavarapu et al., 2017; Barrios et al., 2018; Hurtado et al., 2021].

In this present paper, Long Short-Term Memory deep neural networks were tested to estimate missing monthly precipitation data. The application was done on the K'sob basin in Algeria. LSTM was compared with the most widely used classical methods such as the inverse distance weighting method and the coefficient of correlation weighting method.

METHODOLOGY

Long short-term memory model

Long short-term memory was introduced by [Hochreiter and Schmidhuber, 1997]. It is a powerful architecture of the recurrent neural network (RNN). LSTM is designed to overcome the error backflow problems. LSTM has the ability to perform complex artificial tasks that no other recurrent net algorithm has solved [Hochreiter and Schmidhuber, 1997]. The architecture of the LSTM unit is presented in Figure 1. The LSTM equations are listed below. l

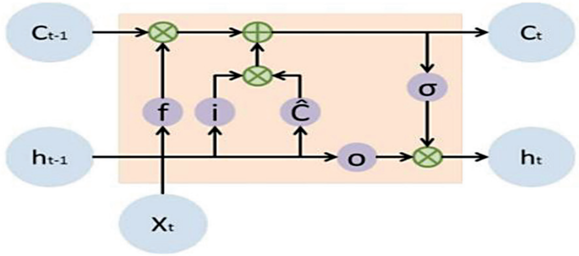

Figure 1. The memoru unit structure fo the LSTM layer [Rahimzad et al., 2021]

$$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i) \qquad (1)$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \qquad (2)$$

$$o_t = \sigma(W_0 x_t + U_0 h_{t-1} + b_0) \qquad (3)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \qquad (4)$$

$$C_t = f_t \otimes C_{t-1} + i_{t-1} \otimes \tilde{C}_t \qquad (5)$$

$$h_t = o_t \tanh \otimes (C_{t-1}) \qquad (6)$$

where: i_t, f_t, and o_t – the input gate, forget gate, and output gate, respectively;

W_i, $W_{f,}$ and W_o – the weights linking the input, forget, and output gates with the input, respectively;

U_i, U_f, and U_o – the weights from the input, forget, and output gates to the hidden layer, respectively;

b_i, b_f, and b_o – the input, forget, and output gate bias vectors, respectively;

$\tilde{C}_t$ – the state of the cell at the previous time; C_t is the current state of the cell; h_{t-1} – the output of the cell at the previous time point; h_t refers to the output of the cell at the current time.

Traditional methods

The LSTM was compared with the most widely used classical methods including the inverse distance weighting method and the coefficient of correlation weighting method.

Inverse distance weighting method

The IDWM is one of the most traditionally used methods for estimating missing precipitation data. The estimation of missing value of an observation, P_m, at a base station m, is calculated using the observed values at other stations and the distance between the base station and the other stations using the following equation:

$$P_m = \frac{\sum_{i=1}^{n} P_i d_{mi}^{-k}}{\sum_{i=1}^{n} d_{mi}^{-k}} \quad (7)$$

where: P_m is the precipitation at the base station m;
n is the number of stations;
P_i is the precipitation at station i,
d_{mi} is the distance between the station m and the station I;
k is referred to as friction distance [Vieux, 2001] that ranges from 1 to 6.
The mostly commonly used value for k is 2.

Coefficient of correlation weighting method

In CCWM, the coefficients of correlation between the data of station m and the other stations are used as weighting factors and the estimation method is given by [Teegavarapu and Chandramouli, 2005]:

$$P_m = \frac{\sum_{i=1}^{n} P_i R_{mi}}{\sum_{i=1}^{n} R_{mi}} \quad (8)$$

where: R_{mi} is the coefficient of correlation between the data of station m and any other station i.

Performance measures

The performance of the proposed estimating methods is evaluated using the most widely used goodness of fit measures.

Nash-Sutcliffe efficiency coefficient

Nash-Sutcliffe efficiency (NSE) is a dimensionless goodness-of-fit indicator, introduced by [Nash and Sutcliffe, 1970]. It has a range from -∞ to1. The value 1 indicates perfect fit, while a NSE ≤ 0 suggests that the mean of the observed values is a better predictor than the model. NSE is given by:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(P_i - \widehat{P_i})^2}{\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad (9)$$

where: P_i is the observed precipitation;
$\widehat{P_i}$ is the estimated precipitation.

Root mean squared error and Mean absolute error

Root mean squared error (RMSE) and Mean absolute error (MAE) are frequently used, which describe the difference between the model simulations and observations in the units of the variable [Legates and McCabe, 1999]. They are given by the following expressions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - \widehat{P_i})^2}{N}} \quad (10)$$

$$MAE = \frac{\sum_{i=1}^{N}\left|P_i - \widehat{P_i}\right|}{N} \quad (11)$$

Description of study area and data set

The case study area is K'sob watershed (Fig. 2), located in the northeast of Algeria; it covers 1,480 km^2, between altitudes 585 m and 1,888 m. It has a semiarid climate with a continental tendency with a relatively wet winter and a dry and hot summer. Average interannual precipitation is 340 mm. Monthly rainfall data from 5 stations are used in the present work, the characteristics of which are presented in Table 1.

RESULTS AND DISCUSSION

The LSTM deep neural networks model and two weighting methods, namely, inverse distance weighting method and the coefficient of correlation weighting method are used to estimate missing monthly rainfall data at the base station (i.e., Medjez). The data at this base station are missing for the purpose of testing the estimation method.During the operating period of the five stations, there are 21 years of concomitant observations that are useable in the present work. The monthly data of the five stations was divided into two parts, the first part for the calibration with 70% (175 months) and the second part for the testing with 30% (77 months).

LSTM model

There is no precise rule for the choice of the number of hidden layers and the corresponding number of hidden nodes. For this reason, in this work, the optimal architecture of LSTM model is adjusted by a trial-and-error-procedure. The number of the hidden layers is varied from 1

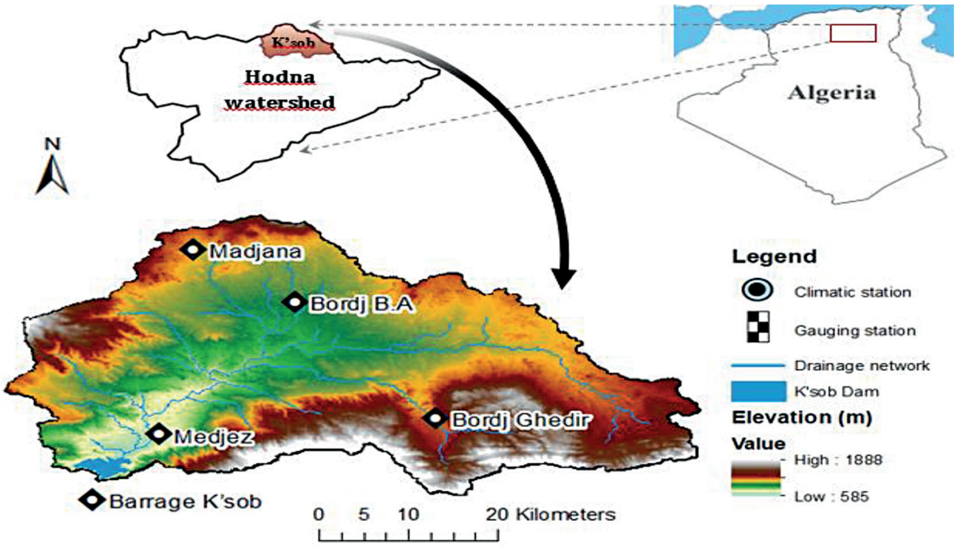

Figure 2. Location of rain gauge stations used in the study

Table 1. Used rain gauge stations

Raingauge stations (code)	Elevation (m)	Geographic coordinates		Statistical properties of monthly rainfall series					
		X	Y	Mean (mm)	Min (mm)	Max (mm)	Standard deviation	Kurstosis	Skewness
Medjez (050901)	684.57	4°37'39.7"	35°53'14.3"	19.3	0.0	108.0	20.51	2.52	1.53
Bordj Ghedir (050904)	1101.83	4°54'23.5"	35°54'27.6"	32.6	0.0	134.2	28.57	0.36	1.00
Bordj B.A. (050905)	880.12	4°45'52.5"	36°3'22"	29.8	0.0	124.2	25.64	1.01	1.14
Madjana (050906)	1051.27	4°39'44.3"	36°7'24.2"	34.1	0.0	231.0	34.93	6.33	2.07
Barrage K'sob (051005)	550.24	4°33'39.9"	35°48'13.4"	19.4	0.0	133.2	20.26	6.38	1.95

to 10. The number of hidden nodes is varied from 1 to 20. In the present paper, the training applies the ADAM algorithm with constant learning rate of 0.05. To avoid overfitting, a dropout layer with dropout probability of 0.5 was considered. Figure 3 presents the NSE and RMSE evolution in the test part of data for 1 hidden layer. From this figure, one can see that the optimal hidden node number is 12 with NSE and RMSE, equal to 0.77 and 10.73 mm,

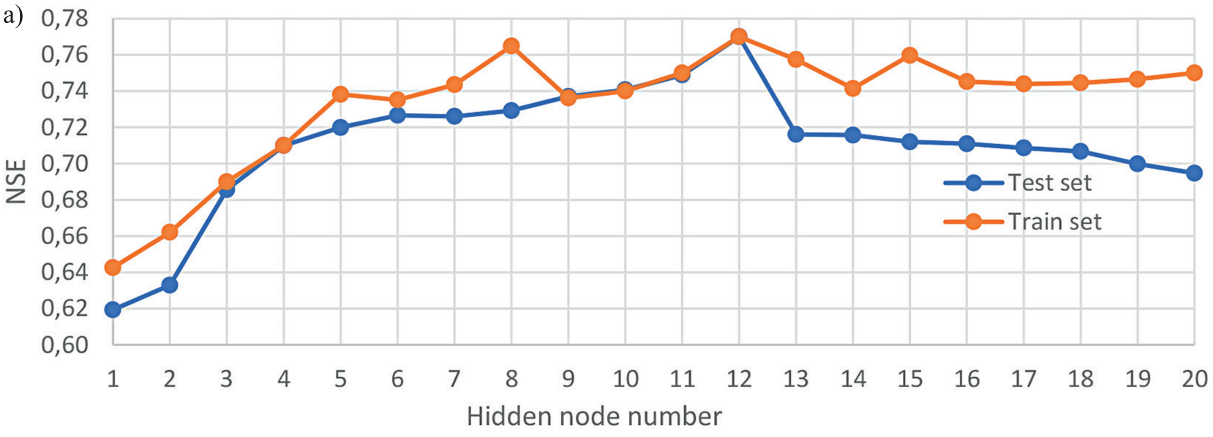

Figure 3. Evolution of a) NSE and b) RMSE for different hidden node number (for 1 hidden layer)

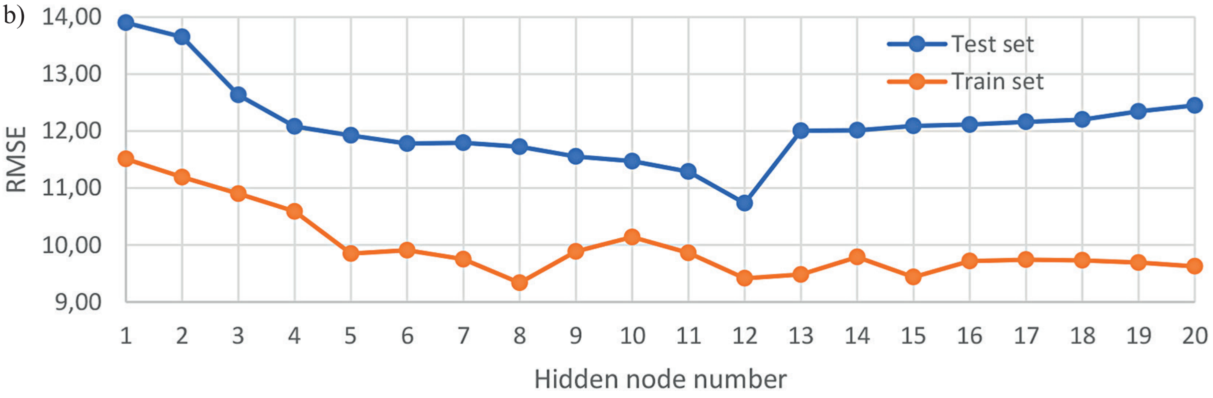

Figure 3. Cont. Evolution of a) NSE and b) RMSE for different hidden node number (for 1 hidden layer)

Table 2. Results from LSTM model

Hidden layer number	Training			Testing			Optimal hidden node number
	NSE	RMSE (mm)	MAE (mm)	NSE	RMSE (mm)	MAE (mm)	
1	0.76	9.42	7.27	0.77	10.73	8.86	12
2	0.84	7.66	5.56	0.81	9.89	7.88	12
3	0.93	4.98	3.48	0.82	9.64	6.69	12
4	0.93	5.15	3.90	0.84	9.06	6.68	14
5	0.96	4.08	3.13	0.83	9.25	6.90	10
6	0.94	4.78	3.31	0.82	9.61	7.03	6
7	0.95	4.45	3.17	0.82	9.57	7.04	12
8	0.93	4.96	3.70	0.82	9.55	7.06	8
9	0.95	4.34	3.51	0.82	9.63	7.26	6
10	0.95	4.24	2.94	0.82	9.63	7.25	6

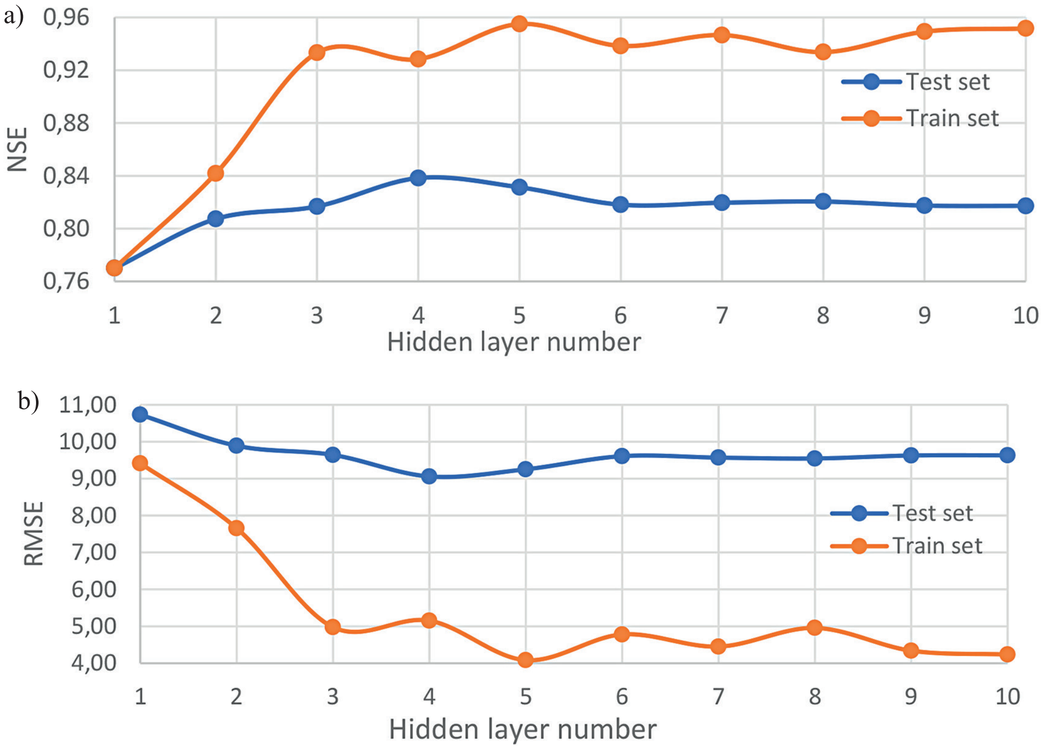

Figure 4. Evolution of a) NSE and b) RMSE for different hidden layer number

respectively. Therefore, this model was chosen as the best alternative for 1 hidden layer. The same procedure was adopted for the other hidden layer number (from 2 to 10). The obtained results are presented in Table 2.

Figure 4 presents the evolution of NSE and RMSE for different hidden layer number. It is possible from this figure that it can be shown as the best LSTM model with 4 hidden layer and 14 hidden units on the bases of 0.84 and 9.06 mm, according to NSE and RMSE, respectively. A comparison between observed and estimated data using the best LSTM model is given in Figure 5.

Figure 5 shows that the LSTM model has good estimation ability, with little time shift error as the extreme events for the estimated values correspond to the extreme values of the observed values. The corresponding scatter plot indicates that the points are closer to the trend line and there are no points of significant overestimation or underestimation.

IDWM and CCWM methods

The application of these two methods consists in calculating the distance between each station

Table 3. Weighting factors for IDWM and CCWM

Station	D (km)	R
Bordj Ghedir (050904)	25.03	0.80
Bordj B.A. (050905)	20.26	0.77
Madjana (050906)	21.40	0.59
Barrage K'sob (051005)	8.70	0.90

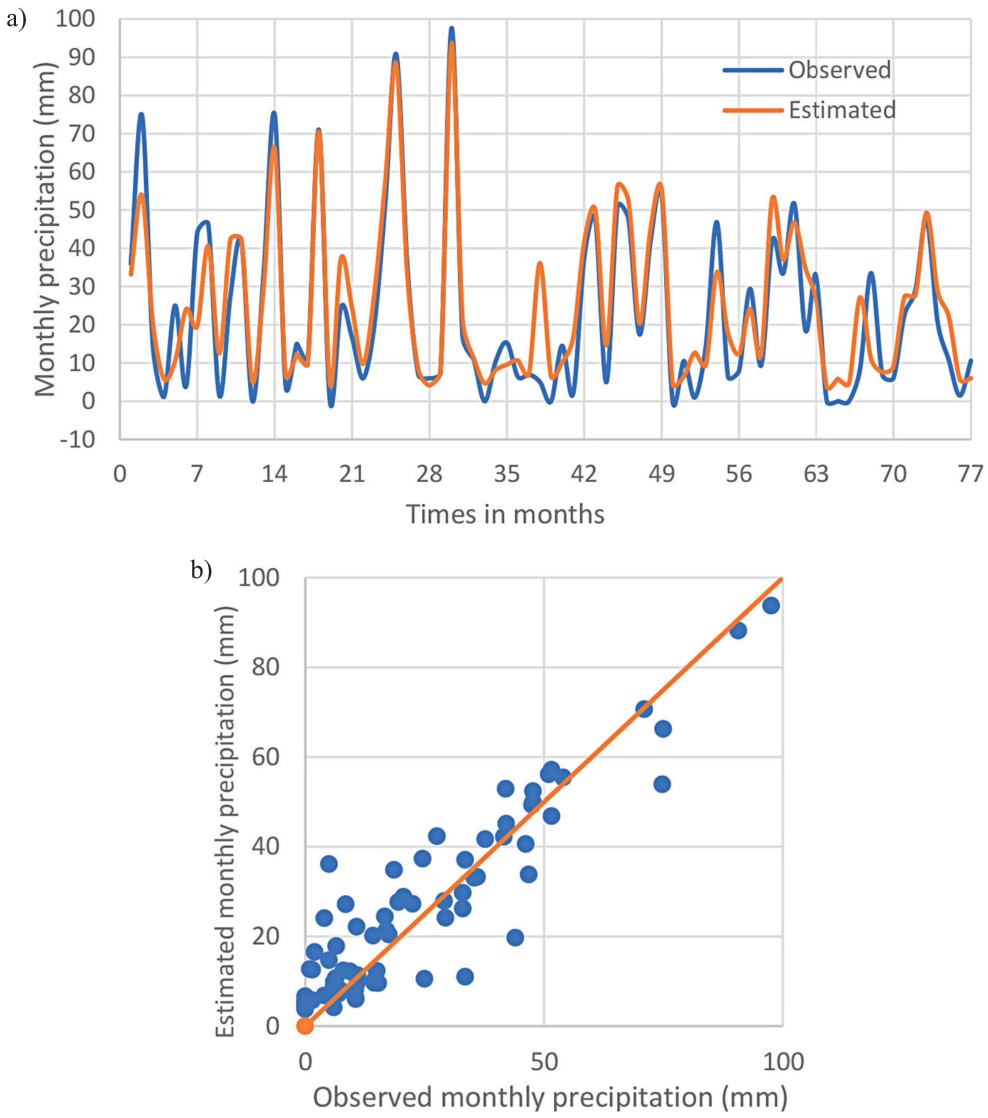

Figure 5. Comparison between observed and estimated data using LSTM: a) Observed and estimated data using LSTM; b) Scatter plot from LSTM

and the base station for the IDWM, and these distances are used as weighting factors to calculate the missing precipitation value (Eq. 7). The final weighting factors are presented in Table 3.

For CCWM, the distance is replaced by the correlation coefficient between the base station and the other stations (Eq. 8). A comparison between observed and estimated data using CCWM and IDWM is given in Figures 6 and 7, respectively.

Figure 6 shows that the CCWM methods has a poor estimation ability compared to the LSTM model. The corresponding scatter plot indicates an overestimation of the precipitation data. Figure 7 is for the IDWM model that has a very poor estimation ability compared to the LSTM model and the CCWM methods with a significant underestimation of precipitation data. Table 4 presents the performance measures of the best obtained LSTM model (CCWM and IDWM).

In addition to the graphical results discussed above, one can see from this table that the LSTM model gives the best performance measures compared to the IDWM and CCWM methods. For IDWM, NSE is equal to -0.11 which is an indication of a very poor model. Compared to IDWM, the CCWM estimation results are slightly improved with NSE equal to 0.37, which always remains at a low NSE value for an acceptable model estimation.

Table 4. Best obtained models

Performance measures	LSTM	CCWM	IDWM
RMSE (mm)	9.06	17.95	23.71
MAE (mm)	6.68	14.21	16.49
NSE	0.84	0.37	-0.11

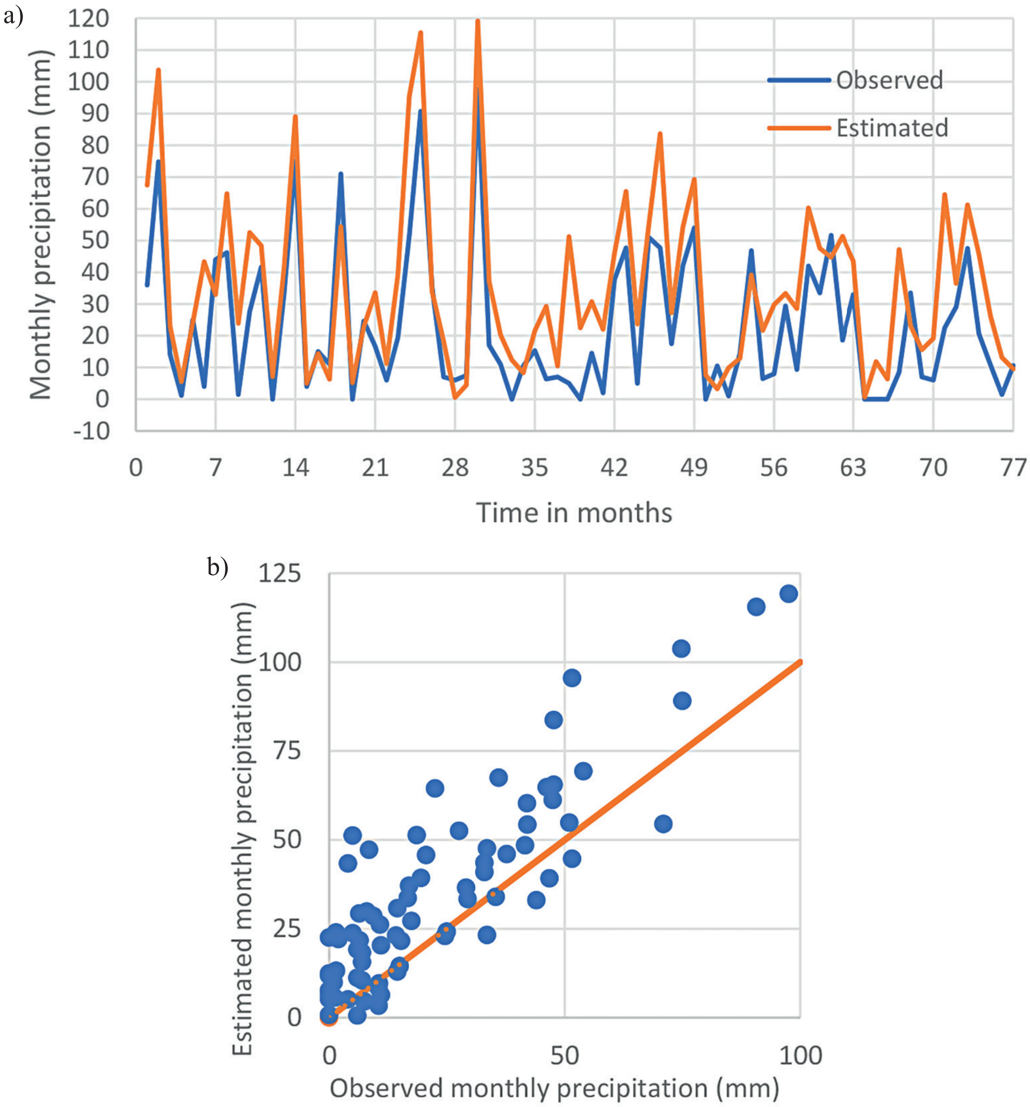

Figure 6. Comparison between observed and estimated data using CCWM: a) Observed and estimated data using CCWM; b) Scatter plot from CCWM

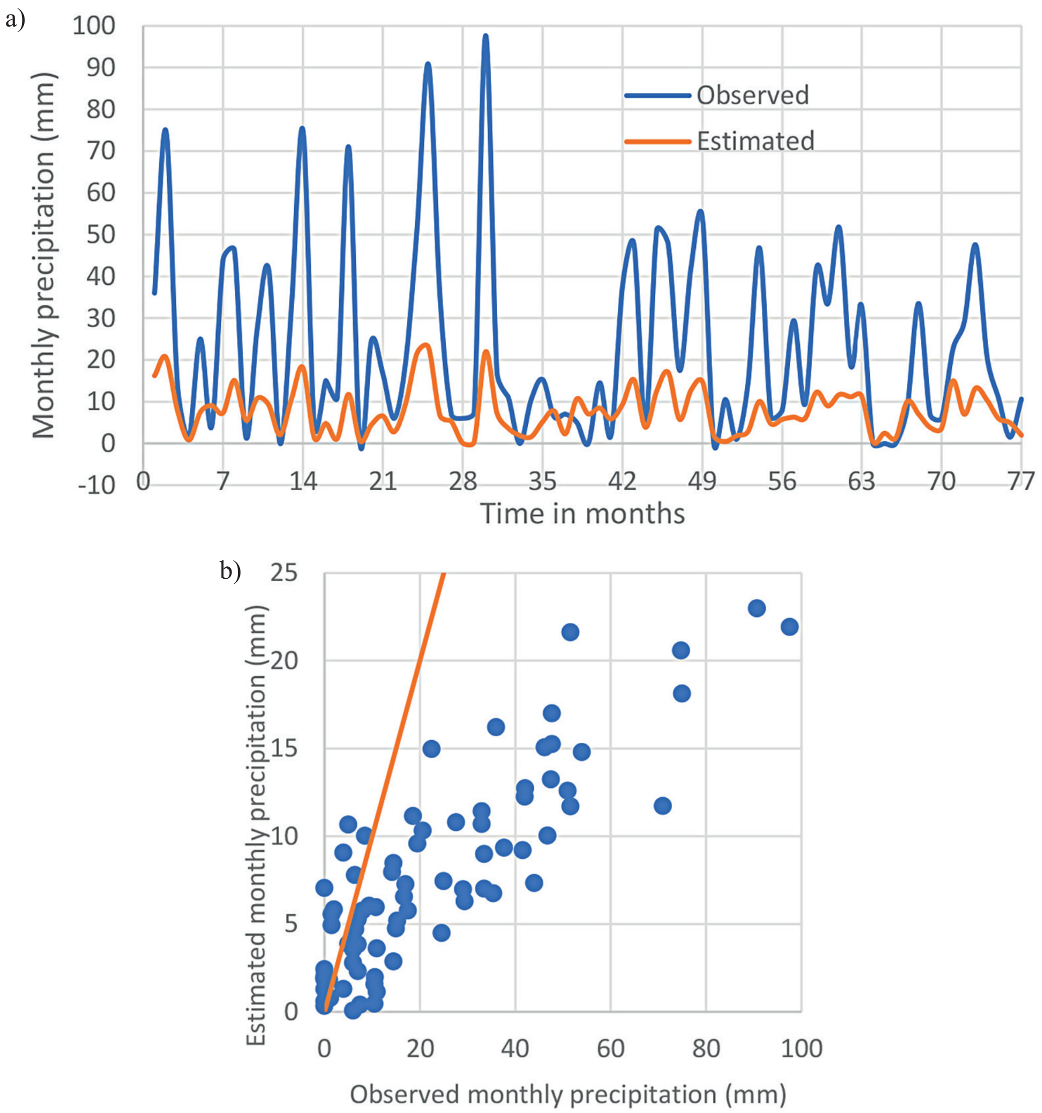

Figure 7. Comparison between observed and estimated data using IDWM: a) Observed and estimated data using IDWM; b) Scatter plot from IDWM

It should be noted that the LSTM model considerably improves the estimation results, compared to those obtained from CCWM and IDWM. For NSE it has been significantly improved from 37% to 84%, also the RMSE from 17.97 to 9.06 mm and the MAE from 14.21 to 6.68 mm, which is an improvement of 50%.

CONCLUSIONS

Accurate estimation of missing precipitation data is crucial for practically all types of hydrological modeling. This paper investigates the accuracy of the Long Short-Term Memory model for estimating missing precipitation monthly data in the K'sob basin in Algeria.

The performance of the LSTM model is compared with the most widely used classical methods including inverse distance weighting method and the coefficient of correlation weighting method. Comparison and evaluation of methodological estimations are based on various performance measures, including the Nash-Sutcliffe model efficiency coefficient, the root Mean Squared Error and the Mean Absolute Error. The following conclusions can be drawn from the study.

- LSTM model performed better than the other methods, with 0.84, 9.06 mm and 6.68 mm according to NSE, RMSE and MAE, respectively.
- IDWM and CCWM gave poor estimation results.
- CCWM performed slightly better than IDWM, which can be explained by the fact that the

correlation coefficient constitutes a better weighting factor than the distance.
- The LSTM network architecture (number of hidden layers and number of hidden nodes) has a considerable influence on the model performance.

It was noted that NSE improved from 0.77 (for 1 hidden layer) to 0.84 (for 4 hidden layers) and from 0.75 (with 2 hidden units) to 0.84 (with 14 hidden units).

On the basis of the obtained results, the LSTM model can be used to estimate the missing precipitation data in all Algerian basins.

REFERENCES

1. Aguilera H., Guardiola-Albert, C., Serrano-Hidalgo C. 2020. Estimating extremely large amounts of missing precipitation data. Journal of Hydroinformatics, 22(3), 578–592.
2. Bárdossy A., Pegram G. 2014. Infilling missing precipitation records – A comparison of a new copula-based method with other techniques. Journal of Hydrology, 519, 1162–1170.
3. Barrios A., Trincado G., Garreaud R.2018. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in South-Central Chile. Forest Ecosystems, 5(1), 28.
4. Coulibaly P., Evora N. D. 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology, 341(1–2), 27–41.
5. Garcia M., Peters-Lidard C. D., Goodrich D. C. 2008. Spatial interpolation of precipitation in a dense gauge network for monsoon storm events in the southwestern United States: MONSOON RAINFALL INTERPOLATION. Water Resources Research, 44(5).
6. Hasanpour Kashani M., Dinpashoh, Y. 2012. Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment, 26(1), 59–71.
7. Hochreiter S., Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8), 1735–1780.
8. Hurtado S.I., Zaninelli P.G., Agosta E.A., Ricetti L. 2021. Infilling methods for monthly precipitation records with poor station network density in Subtropical Argentina. Atmospheric Research, 254, 105482.
9. Kajornrit J., Wong K. W., Fung C. C. 2012. Estimation of missing precipitation records using modular artificial neural networks. International Conference on Neural Information Processing, 52–59.
10. Kim J.-W., Pachepsky Y. A. 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology, 394(3–4), 305–314.
11. Legates D.R., McCabe J.G.J. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1), 233–241.
12. Londhe S., Dixit P., Shah S., Narkhede, S. 2015. Infilling of missing daily rainfall records using artificial neural network. ISH Journal of Hydraulic Engineering, 21(3), 255–264.
13. Ly S., Charles C., Degré A. 2011. Geostatistical interpolation of daily rainfall at catchment scale: The use of several variogram models in the Ourthe and Ambleve catchments, Belgium. Hydrology and Earth System Sciences, 15(7), 2259–2274.
14. Mital U., Dwivedi D., Brown J. B., Faybishenko B., Painter S. L., Steefel C. I. 2020. Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. Frontiers in Water, 2, 20.
15. Nash J.E., Sutcliffe J.V. 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology, 10(3), 282–290.
16. Paulhus J.L.H., Kohler M.A. 1952. Interpolation OF Missing Precipitation Records. Monthly Weather Review, 80(8), 129–133.
17. Rahimzad M., Moghaddam Nia A., Zolfonoon H., SoltaniJ., Danandeh Mehr A., Kwon H.-H. 2021a. Performance Comparison of an LSTM-based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting. Water Resources Management, 35(12), 4167–4187.
18. Shepard D. 1968. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM National Conference, 517–524.
19. Simanton J.R., Osborn H.B. 1980. Reciprocal-distance estimate of point rainfall. Journal of the Hydraulics Division, 106(7), 1242–1246.
20. Suhaila J., Sayang M.D., Jemain A.A. 2008. Revised spatial weighting methods for estimation of missing rainfall data. Asia-Pacific Journal of Atmospheric Sciences, 44(2), 93–104.
21. Teegavarapu R.S. 2007. Use of universal function approximation in variance-dependent surface interpolation method: An application in hydrology. Journal of Hydrology, 332(1–2), 16–29.
22. Teegavarapu R.S. 2012. Spatial interpolation using nonlinear mathematical programming models for estimation of missing precipitation records. Hydrological Sciences Journal, 57(3), 383–406.
23. Teegavarapu R.S., Aly A., Pathak C.S., Ahlquist J., Fuelberg H., Hood J. 2018. Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: Use of optimal weighting parameters and nearest neighbour-based corrections. International Journal of Climatology,

38(2), 776–793.

24. Teegavarapu R.S. Tufail M., Ormsbee L. 2009. Optimal functional forms for estimation of missing precipitation data. Journal of Hydrology, 374(1–2), 106–115.
25. Teegavarapu R.S.V. 2007. Use of universal function approximation in variance-dependent surface interpolation method: An application in hydrology. Journal of Hydrology, 332(1–2), 16–29.
26. Teegavarapu R.S.V. 2014. Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods. Hydrological Sciences Journal, 59(11), 2009–2026.
27. Teegavarapu R.S.V. 2020. Precipitation imputation with probability space-based weighting methods. Journal of Hydrology, 581, 124447.
28. Teegavarapu R.S.V., Chandramouli V. 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 312(1–4), 191–206.
29. Vieux B.E. 2001. Distributed hydrologic modeling using GIS. In Distributed hydrologic modeling using GIS. Springer, 1–17.
30. Wei T.C. 1973. Reciprocal Distance Squared Method, A computer technique for estimating areal precipitation (Vol. 8). US Department of Agriculture, Agricultural Research Service, North Central.
31. Xu W., Zou Y., Zhang, G., Linderman M. 2015. A comparison among spatial interpolation techniques for daily rainfall data in Sichuan Province, China. International Journal of Climatology, 35(10), 2898–2907.